Relational Mining for Compliance Risk

David DeBarr, MITRE and Maury Harwood, IRS

Paper Prepared for
2004 IRS Research Conference
Washington, D.C.
June 2004

Traditional IRS processing has focused on compiling all data into a central set of files for each particular taxpayer, ignoring a rich and growing pool of outside information and the relationships various taxpayers have to each other. NHQ Research funded a proof-of-concept developed by MITRE[i] to test the usefulness of link analysis and relational mining techniques. The data set studied included K-1 data from flow-through entities, as well as the associated business and individual tax return data. The techniques that were investigated included link analysis, graph partitioning, clustering, visualization, graph matching and advanced data mining algorithms. These techniques are complementary in that they reveal different aspects of K-1 networks. Clustering and graph partitioning reveals an overall picture and statistical distribution, while link analysis is useful for reviewing individual networks. Visualization makes it easier to understand networks of a manageable size, less than 200 nodes. Graph matching finds other instances of a graph with particular characteristics, such as possible tax compliance issues.

The proof-of-concept demonstrated compliance with IRS goals and objectives as follows:

- *The ability to identify tax compliance issues in complex K-1 networks. This is especially in regard to corporations and high income individuals that may employ sophisticated schemes and tax shelters to conceal suspicious financial flows.*
- *The ability to identify illegal tax evasion schemes in complex K-1 networks involving distributions to offshore and foreign entities. This applies to illegal schemes organized by tax shelter promoters.*
- *The ability to analyze and understand the characteristics of K-1 networks involving multiple levels of flow through entities.*
- *The ability to identify previously undisclosed abusive tax shelter transactions.*
- *The potential for improvement in tax equity and fairness through analysis of K-1 networks.*

This work showed that the IRS data and domain are well-suited to analysis through graph-based techniques such as graph partitioning and graph-based data mining. However, there is a need for an overall comprehensive strategy and integrated software tools. Current work is focused on assessment of strategic compliance risks and specification of a generic approach to developing tools to identify and quantify these risks on a case-by-case basis. To the extent possible, generic tools are being created.

It is impractical to study taxpayer relationships from the perspective of a single operating division. Many of the partnerships to which Large and Mid-Size Business (LMSB) corporations are related, for example, fall within the jurisdiction of the Small

Business/Self Employed (SB/SE) operating division. In addition, the two operating divisions share many of the compliance risks and ultimately must seek joint approaches to addressing them. Many Tax Exempt and Government Entity (TE/GE) taxpayers are also involved in the flow-through activities. This work is being jointly investigated, and the tools and techniques derived from these expenditures will be used by the entire IRS.

As the number of tax return filings continues to increase from year to year, the IRS would like to use computer-based technology to help perform an initial screening of returns, to detect potential abusive activity and fraud using indicators endorsed by compliance experts. Returns should be ranked for review based on area of compliance expertise, the probability of compliance issue, and the suspected dollar value being sheltered. For example, this allows a partnership return with an 80% probability of having a $10,000,000 compliance issue to be assigned a higher rank than a partnership return with a 90% probability of having a $10,000 compliance issue. This paper discusses the use of computers to perform an initial screening of returns for indicators of compliance issues.

Link Analysis tools and graph-based data mining will allow the IRS to make sense of these voluminous filings. Typically, MITRE's expertise has been devoted to tax system modernization. However, internally, MITRE funds several research projects. Graph-based data mining is an important example of one of these. This project has leveraged the internally funded generic research sponsored by MITRE to build tools that can be applied to the IRS domain. This paper describes some of the advanced algorithmic techniques being tested on the data.

Introduction

Developing a risk-based scoring system begins with an expert-developed list of indicators for known compliance issues. It's important for a knowledge acquisition engineer to be able to connect the indicators to examples of known abuse. The key idea is to find an appropriate balance between having the ability to find known abuse and having the ability to generalize to similar abusive scenarios.

A typical targeting project has several well-defined phases:

- Defining inputs, outputs, and evaluation metrics: Identifying inputs includes selection of available return data to be screened. This can include multiple types of forms to provide contextual information; for example, reviewing closely held flow-through entities in conjunction with the returns of high-income taxpayers. The outputs will typically include a ranked list of returns suspected of having a particular compliance issue; for example, partnership returns indicating the use of a straddle for the exclusive purpose of generating offsetting losses for large capital gains. Evaluation metrics should be defined to allow for measurement of the success of the project; for example, measuring whether a compliance expert agrees with the assessed risk for selected returns.
- Obtaining, exploring, and preprocessing data: This phase typically involves descriptive statistics, visualization, and cluster analysis in order to gain familiarity

with the data to be used for targeting. Preprocessing involves coping with possible transcription issues and normalization requirements for possible analysis techniques.

- Building, validating, and testing screening models: Validation is used to select optimal model parameters to be used for assessing compliance risk, while testing actually provides an estimate of accuracy for each model.
- Deploying risk-analysis models: Successful models should be incorporated into the returns processing cycle and be subjected to annual review for re-evaluating accuracy/performance of the model.

Detecting Abusive Transactions with Support Vector Machines

Computers can be used to perform an initial screening of related tax returns in order to prioritize them for further review by compliance specialists. In fact, computers can be trained to recognize abusive transactions in much the same way you would train another human:

- Identify tax returns involving known examples of abusive transactions
- Pick a few representative examples for training
- Point out fields on the forms that indicate the presence of an abusive transaction
- Ask the human to check new returns for similar indications of abuse

A Support Vector Machine (SVM) is an algorithm for learning from data. Given a set of training examples, an SVM will select a smaller set of representative examples (vectors) to support the task at hand. Support Vector Machines can be used for classification tasks, regression tasks, or density estimation tasks. Given a set of training data containing descriptions for sets of related returns, an SVM can be used to estimate a mapping function to assign categorical labels (classification), numeric values (regression), or estimated probabilities (density estimation). Classification can be used to determine if a set of related returns contains indicators associated with an abusive shelter. Regression can be used to estimate the dollar value associated to an off-shore account. Density estimation can be used to determine how similar a set of related returns is to examples of known abuse.

Like many other statistical and machine learning techniques used for learning from data, a dual-class SVM used for classification requires labeled instances of both compliant and non-compliant returns. By using quadratic programming to solve an optimization problem involving separation of the two classes, an SVM assigns Lagrange multipliers (weights indicating importance) to the returns being analyzed. Those returns with non-zero weights are called support vectors, because they represent important examples to be used for distinguishing compliant returns from non-compliant returns. Unfortunately, when discovering abuse associated with a new type of tax shelter, it can be very time-consuming for a compliance expert to go back and find known compliant returns. A single-class SVM can be used to help alleviate this burden.

A single-class SVM can also be used for classification; i.e., finding returns containing indicators of abuse. By providing only examples of known non-compliant returns, a

single-class SVM can learn to recognize similar activity in other returns. The resulting model (support vectors associated with Lagrange multipliers) can then be used to identify abuse in historical data as well as new filings.

For the purposes of this project, we focused on deriving a set of variables to be used to determine if a high-income taxpayer had set up a flow-through entity solely for the purpose of generating losses to offset large gains from another source. A larger set of variables was transformed into 4 variables by using summation and computing ratios. A parallel coordinates plot of the normalized variables for a sample model is shown in Figure 1. Each vertical bar represents an axis for a variable. For example, a tuple of values representing a set of related entities might contain the values (-0.6504, 0.6504, -0.2809, 0.2738). This set of values is illustrated by the black line that begins at the bottom of the first vertical bar. Each numeric value was divided by the Euclidean norm of the observation (the square root of the sum of the squared values), in order to ensure the SVM software package would converge to a global optimum quickly while preserving the existing ratios between numeric values.

The training data for the model came from 32 abusive transactions. The 2 solid black lines in Figure 1 represent support vectors, while the other 30 dotted gray lines represent the remainder of the training data.

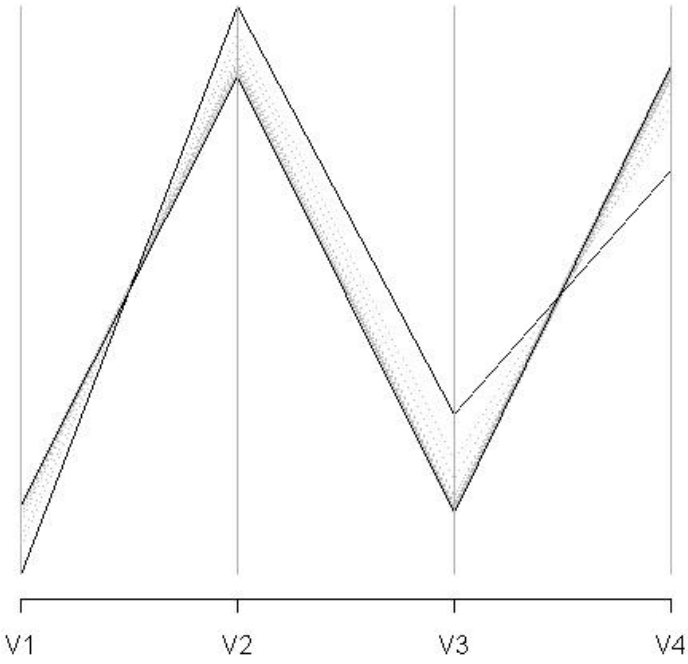

Figure 1 Parallel Coordinates Plot of Normalized Flow-through/Taxpayer Data

A single-class SVM is created by computing Lagrange multipliers (weights indicating importance) for the training examples. The Lagrange multipliers are computed by solving the following constrained optimization problem:

Minimize $\frac{1}{2}\vec{a}^T H \vec{a}$ subject to the following constraints: $0 \le a_i \le \frac{1}{n\,n}$ and $\vec{e}^T \vec{a} = 1$

…where $\vec{a}$ is the vector of Lagrange multipliers to be computed, H is a matrix of numeric outputs of a kernel (similarity) function for the training examples, n is an upper bound on the number of training examples that can be deemed to be outliers (unusual examples), n is the number of training examples, and $\vec{e}$ is just a vector of ones.

The basic idea behind a single-class SVM is to identify similar transactions using a kernel (similarity) function. A Gaussian kernel was selected as the kernel function for this model, and leave-one-out cross-validation was used to find an optimal value for the kernel width parameter. The output model consisted of the two support vectors shown in Figure 1 (the dark lines), with a Lagrange multiplier of 0.48 assigned to each observation[ii]. These cases define the boundaries for the training data. The following discriminant function is used to determine how similar a new observation is to the training data:

$$\sum_{i=1}^{2} 0.48 \exp\left(-14.92 \left\| SupportVector_i - NewObservation \right\|^2\right) - 0.5236$$

The -14.92 coefficient is a kernel width parameter that was found using cross validation, while -0.5236 is a bias term used to adjust the boundary around the training data. A weighted product of the value of the discriminant function and the dollar value of possible abuse is used to rank new sets of returns by compliance risk. For example, during the next tax year, the highest ranked compliance risk involved a $50 million abusive transaction, conducted with the aid of a known promoter. Figure 2 shows a partnership (the ellipse with a thick black border) allocating an offsetting loss (the thick red line) to a high-income taxpayer (the red rectangle with a thick black border) who receives a large capital gain (the thick black line) from another source (the rectangle with a blue border).

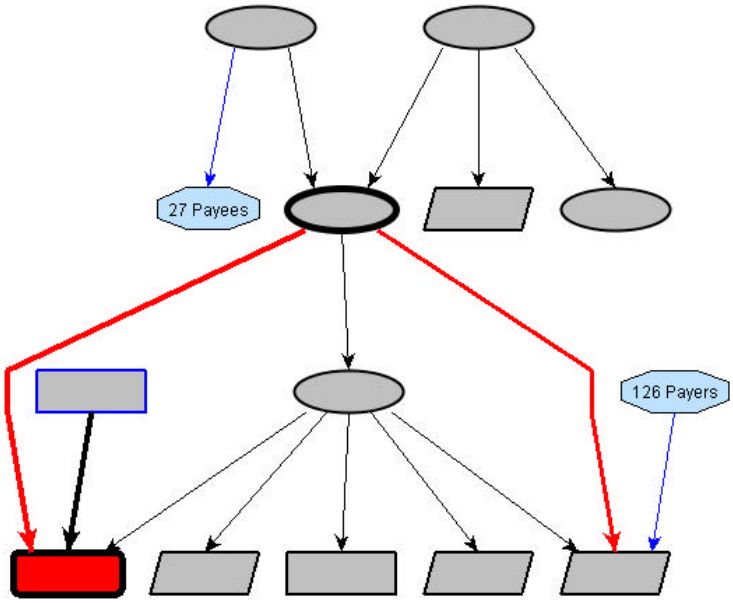

Figure 2 Graph Illustrating an Abusive Transaction Found in the Next Tax Year

While obviously not all classification problems are easy, it appears that some of the most egregious scenarios make it relatively easy to distinguish non-compliant returns from compliant returns. It's not easy to hide a multi-million dollar straddle transaction designed exclusively for the purpose of offsetting large gains. Using data from tax years 2000 to 2003, this model identified numerous abusive transactions, where each transaction involved millions of dollars. These transactions included known violations of Notice 2000-44 and Notice 2002-65; however, several of these transactions do not appear to have been previously discovered by the IRS! The principle advantage of replacing a crisp rule (where either the example meets the selection criteria or it does not) with a similarity function is the ability to rank output by both the likelihood of abuse and the dollar values involved.

Single-class SVMs also enjoy the sound theoretical basis provided by Vladmir Vapnik's Statistical Learning Theory. The core idea is to enable use of small sample statistical inference by accounting for the associated risk appropriately. The optimization problem is cast as Structural Risk Minimization, striking a balance between performance on the training data and bounds on future performance dictated by the amount of training data and the shape/complexity of the decision boundary.

Refining Models with Active Learning

Ranked output from any workload (audit) selection model must be reviewed by a compliance expert to determine if further investigation is warranted. Single-class SVM models can be refined by providing feedback on misclassifications: both misclassifications of non-compliant returns as compliant and misclassifications of compliant returns as non-compliant.

Active learning can help to refine a model quickly, providing increased accuracy with minimal effort. Active learning allows a learning algorithm to select data points for labeling based on the amount of uncertainty associated with the classification of each data point. Using active learning allows the classifier to focus on refining the decision boundary as quickly as possible, whereas further training on randomly selected data points if often unlikely to provide the required information as quickly.

Detecting Promoters by Identifying Unusually Frequent Sets of Values

Promoters of abusive and fraudulent transactions are of special concern. Without promoters, it is unlikely that many taxpayers would use break-even transactions to generate large "paper" losses or off-shore accounts to evade U.S. taxes. Identifying promotions quickly is a useful way to avoid pain and aggravation for both tax payers and tax administrators. One possible method to identify promotions is to look for common connections (values) involved in many abusive transactions. These connections may include a common payee, a shared address, or a shared preparer. Formal statistical tests of independence can be used to identify those values associated with a disproportionate number of possibly abusive transactions compared to the rest of the population. Figure 3 shows substructures employed by a promoter engaged in offshore abuse (sending income to countries offering reduced tax rates with less restrictive reporting requirements).

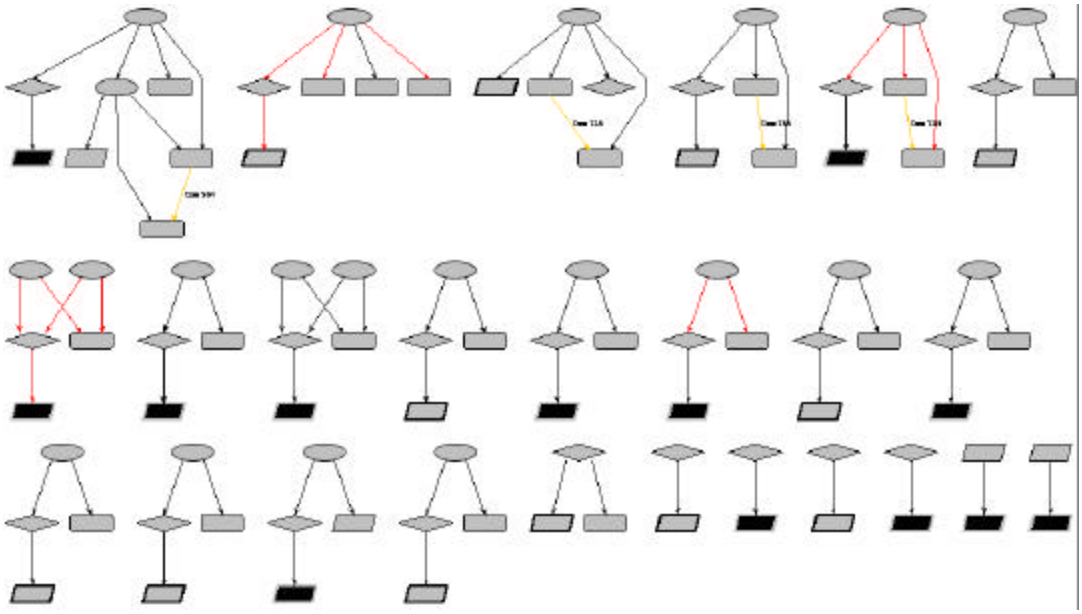

Figure 3 Examples of Abusive Flow-through Structures Involving Off-Shore Entities

All of the on-shore trusts (diamonds) and the off-shore trusts (parallelograms acting as termination points) share a common set of values on their returns. The black nodes indicate off-shores trusts associated with a known tax haven country. There were over 60 entities involved in this scheme. The probability of finding a common set of values for this many trust returns was less than one in a thousand.

Conclusions

It's not possible to have compliance experts review every possible set of related tax returns. In order to meet the strategic goal of "increasing the IRS workforce only slightly while handling an increased workload," computers will need to be used to perform initial screening/ranking of returns for later review by compliance experts. Single-class support vector machines can be used to identify abusive or fraudulent transactions, using only known examples of non-compliance for training data. Active learning can be used to refine targeting models. Common connections between possibly abusive transactions can be used to identify potential promoters of these transactions.

Future Directions

This effort was started in October 2001 with modest funding. Possible activities for the near-term include:

- Deploying the visualization prototype to more selected sites to obtain user feedback on requirements for the visualization capabilities
- Using the existing database of known abusive tax shelters to find out what percentage of these shelters can be identified using algorithms generated from issue specialist inputs
- Asking the issue specialists to evaluate ranked lists of structures identifying possible shelters that have not been previously identified (initial results appear promising; see figures 2 and 3)
- Exploring temporal analysis to identify changes that may also be indicative of abusive behavior.

Acknowledgements

As noted before this work is a collaborative cross business operating division project. As such there have been numerous contributors. Some of these include:

Mike Whalin	Julie Buckel	Kay Wolman
Scott Emerson	Michelle Rhone	George Klima
Susan Kerrick	Dan Killingsworth	Blaine Barkley
John Davidson	Maria Costa	Ann Dario
Larry May	Mike Bland	Don McPartland
Tom Colaiezzi	Jeanette May	Rick Fratanduono
Theresa Hallquist	David Stanley	Michael Israel
Jim Needham	Steve Reed	M C Fusco

Many thanks are owed to a number of groups, including HQ IRS Office of Research Tax Return Database Group, SBSE research, LMSB research, and SBSE issue specialists. We apologize to anyone we may have inadvertently omitted from this list.

Other MITRE collaborators who made significant contributions to this project include Dan Venese, Lowell Rosen, Zohreh Nazeri, and Anne Cady.

References

"IRS Strategic Plan: Fiscal Year 2000 – 2005." Jan 2001.
http://www.irs.gov/pub/irs-utl/irs_strategic_plan.pdf

"Tax Me If You Can." PBS Frontline. Feb 2004.
http://www.pbs.org/wgbh/pages/frontline/shows/tax/shelter/

"Tax Shelters: Who's Buying, Who's Selling, and What's the Government Doing About It?" Senate Finance Committee Hearing. Oct 2003,
http://finance.senate.gov/sitepages/hearing102103.htm

Scholkopf, B, et al. "Estimating the Support of a High-Dimensional Distribution." Neural Computation, 13(7), 2001.

Sung, K. "Learning and Example Selection for Object and Pattern Detection." Ph. D. Thesis. EECS Dept, MIT. Feb 1996.

Tong, S. "Active Learning: Theory and Applications." Ph. D. Thesis. CS Dept, Stanford. Aug 2001.

Vapnik, V. Statistical Learning Theory. John Wiley & Sons. 1998.

[i] *MITRE is the Federally Funded Research and Development Center (FFRDC) for the IRS.*
[ii] For this project, a scaled version of the optimization problem was solved to find the Lagrange multipliers.